JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 6, 2012

Pp. 839–854
DOI: 10.1089/cmb.2012.0029

Normal and Compound Poisson Approximations for Pattern Occurrences in NGS Reads

ZHIYUAN ZHAI,[1] GESINE REINERT,[2] KAI SONG,[3] MICHAEL S. WATERMAN,[4,5] YIHUI LUAN,[1] and FENGZHU SUN[4,5]

ABSTRACT

Next generation sequencing (NGS) technologies are now widely used in many biological studies. In NGS, sequence reads are randomly sampled from the genome sequence of interest. Most computational approaches for NGS data first map the reads to the genome and then analyze the data based on the mapped reads. Since many organisms have unknown genome sequences and many reads cannot be uniquely mapped to the genomes even if the genome sequences are known, alternative analytical methods are needed for the study of NGS data. Here we suggest using word patterns to analyze NGS data. Word pattern counting (the study of the probabilistic distribution of the number of occurrences of word patterns in one or multiple long sequences) has played an important role in molecular sequence analysis. However, no studies are available on the distribution of the number of occurrences of word patterns in NGS reads. In this article, we build probabilistic models for the background sequence and the sampling process of the sequence reads from the genome. Based on the models, we provide normal and compound Poisson approximations for the number of occurrences of word patterns from the sequence reads, with bounds on the approximation error. The main challenge is to consider the randomness in generating the long background sequence, as well as in the sampling of the reads using NGS. We show the accuracy of these approximations under a variety of conditions for different patterns with various characteristics. Under realistic assumptions, the compound Poisson approximation seems to outperform the normal approximation in most situations. These approximate distributions can be used to evaluate the statistical significance of the occurrence of patterns from NGS data. The theory and the computational algorithm for calculating the approximate distributions are then used to analyze ChIP-Seq data using transcription factor GABP. Software is available online (www-rcf.usc.edu/~fsun/Programs/NGS_motif_power/NGS_motif_power.html). In addition, Supplementary Material can be found online (www.liebertonline.com/cmb).

Key words: algorithms, genome analysis, HMM, next generation sequencing, statistical models.

[1]School of Mathematics, Shandong University, Jinan, Shandong, China.
[2]Department of Statistics, University of Oxford, Oxford, United Kingdom.
[3]School of Mathematics, Peking University, Beijing, China.
[4]Molecular and Computational Biology, University of Southern California, Los Angeles, California.
[5]TNLIST/Department of Automation, Tsinghua University, Beijing, China.

1. INTRODUCTION

The study of the occurrences of word patterns in sequences has played an important role in molecular sequence analysis. Here, we shall use *word pattern of length k* and *k-tuple* interchangeably; often word patterns are also just referred to as *words*. For a given k, the frequencies of word patterns of length k form a vector, referred to as sequence signature (Campbell et al., 1999). Sequence signatures of genomic sequences of varying characteristics are usually different. For example, coding and non-coding sequences usually have different signatures and thus sequence signatures can be useful features to distinguish coding from non-coding sequences (Uberbacher and Mural, 1991). Sequence signatures within different parts of a genome tend to be similar, while they differ significantly between genomes (Karlin and Mrázek, 1997, Nekrutenko and Li, 2000). Thus, sequence signatures have been used to study the evolutionary relationship between different genomic sequences (Jun et al., 2010, Karlin and Mrázek, 1997, Sims et al., 2009, Wu et al., 2009), to study horizontal gene transfer (Dalevi et al., 2006, Dufraigne et al., 2005), and to bin sequence reads from metagenomic studies so that reads in the same bin tend to have similar sequence signatures (McHardy et al., 2006). The sequence signatures can also be employed to detect enrichment for short words. For example, the upstream regions of co-regulated genes usually share common transcription factor binding sites (TFBS) referred to as motifs, and thus motifs are usually enriched within these sequences. Finding enriched word patterns within these sequences is a powerful tool for the identification of TFBS (Pavesi et al., 2004).

Due to the many applications of sequence signatures, extensive studies have been carried out to study the distribution of the number of occurrences of word patterns in one or multiple long sequences consisting of independent identically distributed (i.i.d.) letters and sequences generated by both Markov and hidden Markov models (HMM). Several excellent reviews (Reinert et al., 2000, 2005, Schbath, 2000, Schbath and Robin, 2009) and a book (Robin et al., 2005) on this topic are available. The distribution of the number of occurrences of a pattern can also be studied using the so-called ''imbedded Markov chain'' techniques (Kleffe and Langbecker, 1990, Nuel, 2006, Shan and Zheng, 2009). However, the computation of p-values using these techniques can be very time consuming and impractical for long sequences. We recently studied the power of detecting enriched patterns when motifs are randomly distributed along the genome using HMM (Zhai et al., 2010).

In all these studies, one or several long sequences are available and the word pattern occurrences along these long sequences are studied. Rather than providing a few long sequences, recent developments in sequencing technologies make it possible to sequence a large number of relatively short reads (e.g., 30–80 bp for Illumina/Solexa and 300–500 bp for Roche 454) efficiently and economically. These new sequencing technologies have revolutionized current studies of many biological problems including locating genomic regions of TFBS, histone modification, and chromatin structure using ChIP-Seq, resequencing of known genomes for the identification of genetic polymorphisms, and sequencing of unknown genomes. For the applications of these NGS technologies, see recent reviews (Maclean et al., 2009, Mardis, 2008a,b). Although many computational methods have been developed to analyze NGS data, to our knowledge no studies on the distribution of the number of occurrences of word patterns in the sequence reads generated from NGS have been carried out. In this article, we fill this gap. The main challenge compared to word counts in sequences is that, in NGS, two random processes are involved, namely not only the the randomness in the background genome sequence but also the random sampling of the reads from the background sequence.

The study of the distribution of the numbers of occurrences of word patterns from NGS read data has several important applications, in particular, when the complete genome sequences are not available. First, such distributions are important for the comparison of genomes when NGS short reads are available for each genome (Song et al., 2012). Second, they can be used to identify enriched or depleted patterns in genomes whose complete genomes are not known. Such enriched or depleted patterns can be used to characterize the genome sequences. Third, the null distributions of the numbers of occurrences of patterns can be used to identify enriched patterns in ChIP-Seq experiments and such enriched patterns can be useful for the identification of TFBS.

In this article, we not only study the distributions of the numbers of occurrences of word patterns from NGS read data under a suitable null model, but we also address the issue of the power of the count statistics against an alternative model which assumes that there are motifs present in the sequence. Our methodology builds on Zhai et al. (2010), but differs from that article in the consideration of NGS data and the

consideration of both strands of the genome sequences. In the study of word patterns for long sequences, both strands are rarely considered except in Pape et al. (2008). For NGS, the consideration of both strands are essential since the reads can come from both the forward strand and the reverse strand of the genome sequences. We provide simpler approximate distribution for the number of occurrences of word patterns for NGS data than the approximations given in Pape et al. (2008).

The article is organized as follows. In Section 2, we first present the probability models for the background sequence and the sampling process of reads using NGS. Then the results for normal and compound Poisson approximations for the number of occurrences of patterns in NGS reads are given. As the approximations assume that both the length of the reads and the length of the background sequence go to infinity, whereas in reality they are reasonably short, we also give bounds on the approximation errors. We consider both single strand and double strand models. This section forms the core theoretical results of the article. In Section 3, we first present simulation results to show the validity of the theoretical results for both common and rare patterns, and then use the theoretical results to analyze a ChIP-Seq data set from Valouev et al. (2008). It is surprising to see that, even in the control data, some TFBS signals can be identified, indicating that some residue ChIP effects are present in the control data. Using ChIP-Seq data, we are able to identify the consensus patterns of the motif of interest. The article concludes with some discussion on the limitations of the approach and future research directions. Many of the proofs are given in Supplementary Materials (available online at www .liebertonline.com/cmb).

2. METHODS

2.1. Probabilistic models for the background sequence and sampling of sequence reads using NGS

In NGS, a large number of M reads are randomly sampled from the genome. For studying the distribution of the number of occurrences of patterns among the M reads, two random processes are involved. The first randomness comes from the generation of the background genome sequence and the second randomness comes from random sampling of the reads from the background sequence.

As in previous studies reviewed in Robin et al. (2005) and Schbath and Robin (2009), the background sequence is modeled as a homogeneous ergodic Markov chain taking states in the set $\mathcal{A}=\{0, 1, \cdots, L-1\}$ with transition probability matrix $T = (t_{ll'})_{L\times L}$. The Markov chain has a unique stationary distribution π_0. The results in this paper can also be extended to sequences generated by hidden Markov models without too much difficulty.

Next, we model the sampling of reads along the genome sequence using NGS. As it was shown in Zhang et al. (2008) that the homogeneous Lander-Waterman model (Lander and Waterman, 1988) for genomic mapping does not model the read distributions along the genome well, we use a modified version of the Lander-Waterman model to describe the distribution of reads along the genome. We assume that the sampled reads have the same length of β bp. A total of M reads are independently sampled from the genome of length n bp. Each read starts at position i with probability λ_i, $1\le i\le \bar{n}$, where $\lambda_i\ge 0$, $\sum_{i=1}^{\bar{n}}\lambda_i=1$, with $\bar{n}=n-\beta+1$.

Let $\mathbf{w}=\mathbf{w}_1\mathbf{w}_2\cdots\mathbf{w}_w$ be any word pattern of length w with $\mathbf{w}_j\in\mathcal{A}, j=1, 2, \cdots, w$. Then $P(\mathbf{w})=\pi_{\mathbf{w}_1}\prod_{i=1}^{w-1}t_{\mathbf{w}_i\mathbf{w}_{i+1}}$ is the probability of $\mathbf{w}$. Let $\mathcal{N}_{\mathbf{w}}(M, n, \beta)$ be the number of occurrences of $\mathbf{w}$ in these M reads. To calculate the mean of $\mathcal{N}_{\mathbf{w}}(M, n, \beta)$, note first that in each read of length β, the expected number of occurrences of $\mathbf{w}$ is $(\beta - w + 1)P(\mathbf{w})$. As there are M reads, we obtain that

$$\mathbb{E}(\mathcal{N}_{\mathbf{w}}(M, n, \beta))=M(\beta-w+1)P(\mathbf{w}). \tag{1}$$

We study the approximate distribution of $\mathcal{N}_{\mathbf{w}}(M, n, \beta)$ and the approximate joint distribution of $(\mathcal{N}_{\mathbf{w}}(M, n, \beta), \mathbf{w}\in\mathbb{S})$, where $\mathbb{S}$ indicates the set of word patterns. We consider both single strand and double strand models. In the single strand model, we assume that the reads just come from one strand. In the double strand model, the reads can come from either strand of the genome. We allow for the occurrences to overlap. For example if the sequence is 5′-CAATAATATAATAG-3′ and the word is ATA, then we count four occurrences in the single strand model. A clump of pattern $\mathbf{w}$ is a consecutive region of the sequence with overlapping occurrences of $\mathbf{w}$. For the example given above, there are three

clumps of occurrences, one clump (ATATA) of size two and two clumps of size one each. Counting the occurrences of ATA in the complementary sequence 5′-CTATTATATTATTG-3′ also, there are 4 + 1 = 5 occurrences of **w** in the double strand model. Note that we always count from the 5′ end to the 3′ end of the sequences.

2.2. Normal approximation for the number of occurrences of frequent patterns in randomly sampled NGS reads

In this subsection, we present our results for calculating the covariance of $\mathcal{N}_{\mathbf{u}}(M, n, \beta)$ and $\mathcal{N}_{\mathbf{v}}(M, n, \beta)$ under the models described in Subsection 2.1 for any two word patterns **u** and **v**. Proposition 2.1 presents a formula to calculate $E(\mathcal{N}_{\mathbf{u}}(M, n, \beta)\mathcal{N}_{\mathbf{v}}(M, n, \beta))$. The covariance can then be derived using Equation (1). While the covariance of word counts for a single sequence read can be found in Waterman (1995), Proposition 12.1, the following Proposition 2.1 takes the randomness in the starting positions of the sequence reads into account.

Proposition 2.1. *Let $O_1O_2\cdots O_n$ be the underlying sequence of length n. Let* **u** *and* **v** *be two word patterns of length u and v, respectively, with $u \le v$. Assume that $\beta \ge u + v$. Randomly choose M reads of length β from a genome of length n base pairs according to the model in Subsection 2.1 and let $\mathcal{N}_{\mathbf{u}}(M, n, \beta)$ and $\mathcal{N}_{\mathbf{v}}(M, n, \beta)$ be the numbers of occurrences of word patterns* **u** *and* **v** *in these reads, respectively. Then $\mathbb{E}(\mathcal{N}_{\mathbf{u}}(M, n, \beta)\mathcal{N}_{\mathbf{v}}(M, n, \beta))$ can be calculated by*

$$\left(M + M(M-1)\sum_{i=1}^{n-\beta+1}\lambda_i^2\right)\mathbb{E}_{\beta,0}(\mathbf{u},\mathbf{v}) + M(M-1)\sum_{i=1}^{n-\beta+1}\lambda_i\sum_{\eta=1}^{n-i-\beta+1}\lambda_{i+\eta}(\mathbb{E}_{\beta,\eta}(\mathbf{u},\mathbf{v}) + \mathbb{E}_{\beta,\eta}(\mathbf{v},\mathbf{u})),$$

where $\mathbb{E}_{\beta,\eta}(\mathbf{u},\mathbf{v}) = \mathbb{E}(N_{\mathbf{u}}[1,\beta]N_{\mathbf{v}}[\eta+1,\eta+\beta])$, and $N_{\mathbf{w}}[i, i+\beta-1]$ the number of occurrences of word pattern **w** *in $O_iO_{i+1}\cdots O_{i+\beta-1}$.*

Formulas for calculating $\mathbb{E}_{\beta,\eta}(\mathbf{u},\mathbf{v})$ are given in the supplementary materials, Proposition A.1; they are based on a slight modification of the proof for Proposition 12.1. in Waterman (1995).

Proof of Proposition 2.1. Let $C_{\mathbf{w}}(m)$ be the number of occurrences of word pattern **w** in the m-th read, $m = 1, 2, \cdots, M$. Then

$$\mathcal{N}_{\mathbf{w}}(M, n, \beta) = \sum_{m=1}^{M} C_{\mathbf{w}}(m).$$

Let

$$\mathbb{E}_{\beta,\eta}(\mathbf{u},\mathbf{v}) = \mathbb{E}(N_{\mathbf{u}}[1,\beta]N_{\mathbf{v}}[1+\eta,\beta+\eta]).$$

According to our model, it is easy to see that for word patterns **u** and **v**, the counts $(C_{\mathbf{u}}(m), C_{\mathbf{v}}(m))$ have the same distribution for all $m = 1, 2, \cdots, M$. Similarly, for any $m \ne m'$, $(C_{\mathbf{u}}(m), C_{\mathbf{v}}(m'))$ have the same distribution. Thus,

$$\mathbb{E}(C_{\mathbf{u}}(m)C_{\mathbf{v}}(m)) = \mathbb{E}(C_{\mathbf{u}}(1)C_{\mathbf{v}}(1)), \quad \mathbb{E}(C_{\mathbf{u}}(m)C_{\mathbf{v}}(m')) = \mathbb{E}(C_{\mathbf{u}}(1)C_{\mathbf{v}}(2)), \quad m \ne m',$$

and

$$\mathbb{E}(\mathcal{N}_{\mathbf{u}}(M, n, \beta)\mathcal{N}_{\mathbf{v}}(M, n, \beta)) = M\mathbb{E}(C_{\mathbf{u}}(1)C_{\mathbf{v}}(1)) + M(M-1)\mathbb{E}(C_{\mathbf{u}}(1)C_{\mathbf{v}}(2)).$$

Since the Markovian sequence is stationary, $C_{\mathbf{u}}(1)$ has the same distribution as $N_{\mathbf{u}}[1,\beta]$. Thus,

$$\mathbb{E}(C_{\mathbf{u}}(1)C_{\mathbf{v}}(1)) = \mathbb{E}(N_{\mathbf{u}}[1,\beta]N_{\mathbf{v}}[1,\beta]) = \mathbb{E}_{\beta,0}(\mathbf{u},\mathbf{v})$$

Conditioning on the locations of the first and second reads, we have

$$\mathbb{E}(C_{\mathbf{u}}(1)C_{\mathbf{v}}(2)) = \mathbb{E}\left[\left(\sum_{i=1}^{n-\beta+1}\lambda_i N_{\mathbf{u}}[i, i+\beta-1]\right)\left(\sum_{j=1}^{n-\beta+1}\lambda_j N_{\mathbf{v}}[j, j+\beta-1]\right)\right]$$
$$= \sum_{i=1}^{n-\beta+1}\sum_{j=1}^{n-\beta+1}\lambda_i\lambda_j\mathbb{E}(N_{\mathbf{u}}[i, i+\beta-1]N_{\mathbf{v}}[j, j+\beta-1])$$
$$= \mathbb{E}_{\beta,0}(\mathbf{u}, \mathbf{v})\sum_{i=1}^{n-\beta+1}\lambda_i^2 + \sum_{i=1}^{n-\beta+1}\lambda_i\sum_{\eta=1}^{n-i-\beta+1}\lambda_{i+\eta}(\mathbb{E}_{\beta,\eta}(\mathbf{u}, \mathbf{v}) + \mathbb{E}_{\beta,\eta}(\mathbf{v}, \mathbf{u})).$$

The proposition is proved. ■

For the special case that the background sequence is i.i.d., we have the following corollary.

Corollary 2.1. *Suppose that the background sequence is i.i.d. With the same notation as in Proposition 2.1, we have*

1. The covariance of $\mathcal{N}_{\mathbf{u}}(M, n, \beta)$ *and* $\mathcal{N}_{\mathbf{v}}(M, n, \beta)$ *can be calculated as*

$$\left(M + M(M-1)\sum_{i=1}^{n-\beta+1}\lambda_i^2\right)(\mathbb{E}_{\beta,0}(\mathbf{u}, \mathbf{v}) - (\beta-u+1)(\beta-v+1)P(\mathbf{u})P(\mathbf{v}))$$
$$+ M(M-1)\sum_{\eta=1}^{\beta-1}\sum_{i=1}^{n-\beta-\eta+1}\lambda_i\lambda_{i+\eta}(\mathbb{E}_{\beta,\eta}(\mathbf{u}, \mathbf{v}) + \mathbb{E}_{\beta,\eta}(\mathbf{v}, \mathbf{u}) - 2(\beta-u+1)(\beta-v+1)P(\mathbf{u})P(\mathbf{v})).$$

2. If $\lim_{n\to\infty}(n-\beta-\eta+1)\sum_{i=1}^{n-\beta-\eta+1}\lambda_i\lambda_{i+\eta} = r_\eta$ *and* M *depends on* n *such that* $\lim_{n\to\infty} M/n = \theta$, *then*

$$\lim_{n\to\infty}\frac{\mathrm{Cov}(\mathcal{N}_{\mathbf{u}}(M, n, \beta), \mathcal{N}_{\mathbf{v}}(M, n, \beta))}{M} = (1+\theta r_0)(\mathbb{E}_{\beta,0}(\mathbf{u}, \mathbf{v}) - (\beta-u+1)(\beta-v+1)P(\mathbf{u})P(\mathbf{v}))$$
$$+ \theta\sum_{\eta=1}^{\beta-1} r_\eta(\mathbb{E}_{\beta,\eta}(\mathbf{u}, \mathbf{v}) + \mathbb{E}_{\beta,\eta}(\mathbf{v}, \mathbf{u}) - 2(\beta-u+1)(\beta-v+1)P(\mathbf{u})P(\mathbf{v})).$$

In particular, if the reads are uniformly sampled from the genomic sequence, i.e. $\lambda_i = 1/(n-\beta+1)$, *then* $r_\eta = 1, \eta = 1, 2, \cdots, \beta-1$.

Corollary 2.1 follows by noting that for the i.i.d. case and $\eta \geq \beta$,

$$\mathbb{E}_{\beta,\eta}(\mathbf{u}, \mathbf{v}) = \mathbb{E}(N_{\mathbf{u}}[1, \beta])\mathbb{E}(N_{\mathbf{v}}[\eta+1, \eta+\beta]) = \mathbb{E}(N_{\mathbf{u}}[1, \beta])\mathbb{E}(N_{\mathbf{v}}[1, \beta]) = (\beta-u+1)(\beta-v+1)P^2(\mathbf{w}).$$

The second part follows directly by taking the limit of $\mathrm{Cov}(\mathcal{N}_{\mathbf{u}}(M, n, \beta), \mathcal{N}_{\mathbf{v}}(M, n, \beta))$ in Part 1) over M as n tends to infinity.

Given the approximate mean and variance of $\mathcal{N}_{\mathbf{w}}(M, n, \beta)$, it is tempting to approximate the distributions of $\mathcal{N}_{\mathbf{w}}(M, n, \beta)$ using a normal distribution. The approximation is based on the heuristic that the counts in different reads are independent unless the reads overlap, and if the words are not too long, the count in each read would be approximately normally distributed.

As reads are not very long, the approximation error may not be negligible, and hence we give an upper bound on the approximation error. Our result is phrased in terms of

$$d_K(\text{standardized count}, Z) = \sup_x |\mathbb{P}(\text{standardised count } \leq x) - \mathbb{P}(Z \leq x)|,$$

where Z denotes a standard normal variable. Thus $\mathbb{P}$ (standardized count $\leq x$) $\leq \mathbb{P}(Z \leq x) + d_K$, and a bound on d_K can be used to obtain a conservative p-value for the observed standardized count.

Here, we employ Theorem 2.6 in Chen and Shao (2004), and assume the i.i.d. model for the underlying background sequence. Then $N_{\mathbf{w}}[i, i+\beta-1]$ and $N_{\mathbf{w}}[j, j+\beta-1]$ are independent unless $|i-\mathrm{j}| \leq \beta$, where $N_{\mathbf{w}}[l, l']$ is the number of occurrences of word $\mathbf{w}$ in the interval $[l, l']$ along the long sequence. Using the notation $\sigma = \sqrt{\mathrm{Var}(\mathcal{N}_{\mathbf{w}}(M, n, \beta))}$ and $\bar{n} = n-w+1$, the following result holds.

Theorem 2.1. *Assume the i.i.d. model for the background sequence and let Z be standard normally distributed. Then for a word* **w** *of length w,*

$$d_K\left(\frac{1}{\sigma}(\mathcal{N}_{\mathbf{w}}(M,n,\beta)-M\bar{\beta}P(\mathbf{w})),Z\right)$$
$$\leq 2M\sum_{i=1}^{\bar{n}}\lambda_i^2+375(10\bar{\beta}+1)^2\frac{1}{\sigma^3}\sum_{i=1}^{\bar{n}}\{M\bar{\beta}P(\mathbf{w})\lambda_i\}^{\frac{3}{4}}(945\bar{\beta}^2wP(\mathbf{w})\max(1,(M\lambda_i)^3)+4(\bar{\beta}P(\mathbf{w}))^3)^{\frac{3}{4}},$$

where $\bar{\beta}=\beta-w+1$.

In the case that all letters are equally likely and independent, and all $\lambda_i=\bar{n}^{-1}$, the bound will be of order $O((\ln n)^{-1})$ when the word length is not too large, $w < \log_L(\beta/\ln(n))$, while the read length $\beta = c_1L^w \ln n$ for a constant $c_1 > 1$ and the number of reads $M = c_2n/(\ln n)$ for a constant $c_2 > 0$. This type of regime is rather specific, for example the above regime with $n = 5{,}000$ and $w = 4$ on a 4-letter alphabet would require $\beta > 2{,}181$, while $n/(\ln n) = 587$; if $n = 20{,}000$ and $w = 7$ we would need $\beta > 162,259$, while $n/(\ln n) = 2019.5$. Moreover, the above regime would require that M/n is small. Hence, it is no surprise that the normal approximation does not work well in many situations.

In particular, in many practical applications of NGS, the coverage of the sequenced reads is moderate to high depending on the biological applications. Thus, the normal approximation may not work well in these situations. The theorems also explain the poor performance of normal approximation in our simulations in Section 3. We emphasize that the bound may not be the best possible in all settings.

A similar result is available for multivariate word counts. The generalization to a Markovian sequence is straightforward, using the arguments from Huang (2002) for the joint counts starting at a specified position, and a local dependence argument as above.

Finally, note that if $M\lambda_i$ is large for some i, then $\mathcal{N}_{\mathbf{w}}(M,n,\beta)$ might be better approximated by the sum of products of normally distributed variables, which is approximately normal only when the number of summands is large.

2.3. Compound Poisson approximation for the number of occurrences of rare patterns in randomly sampled NGS reads

For rare patterns along the background sequence, the normal approximation as described in Subsection 2.2 is not appropriate; instead, we present a compound Poisson approximation for the number of occurrences of such patterns. This compound Poisson approximation takes clumps of occurrences directly into account. Recall that a clump of word **w** is a maximum consecutive region of the background sequence with overlapping occurrences of **w**. For the clumps, we first introduce some notation. Let $\mathbf{w}^{(p)}=\mathbf{w}_1\mathbf{w}_2\cdots\mathbf{w}_p$ be the p-th prefix composed of the first p letters of **w**. The set of periods of the word **w** of length **w** (Guibas and Odlyzko, 1981, Lothaire, 1983) is defined by

$$\mathcal{P}(\mathbf{w})=\{p\in\{1,2,\cdots,w-1\}:\mathbf{w}_i=\mathbf{w}_{i+p},\text{ for any } i=1,2,\cdots,w-p\}.$$

The set of principal periods of word pattern **w**, $\mathcal{P}'(\mathbf{w})$, are those periods that cannot be written as multiples of other periods. It was shown in Reinert and Schbath (1998) and Schbath (1995) that the number of clumps, $N_{c,n}$, in a Markovian sequence of length n can be approximated by a Poisson random variable with mean $\Lambda(\mathbf{w})=(n-w+1)\hat{\mu}(\mathbf{w})$, where

$$\hat{\mu}(\mathbf{w})=P(\mathbf{w})-\sum_{p\in\mathcal{P}'(\mathbf{w})}P(\mathbf{w}^{(p)}\mathbf{w}). \tag{2}$$

We also consider X_i, the number of occurrences of word pattern **w** in the i-th clump. Let

$$C_k=\{\mathbf{w}^{(p_1)}\mathbf{w}^{(p_2)}\cdots\mathbf{w}^{(p_{k-1})}\mathbf{w}:\ p_1,p_2,\cdots p_{k-1}\text{ are the principal periods of }\mathbf{w}\}$$

be the set of all possible ways a clump of size k can occur. It was shown in Reinert and Schbath (1998) and Schbath (1995) that

$$\mathbb{P}(X_i=k)=\hat{\mu}_k(\mathbf{w})/\hat{\mu}(\mathbf{w}), \tag{3}$$

where

$$\hat{\mu}_k(\mathbf{w}) = P(C_k) - 2P(C_{k+1}) + P(C_{k+2}), \quad k = 1, 2, \cdots. \tag{4}$$

A compound Poisson approximation for $\mathcal{N}_{\mathbf{w}}(M, n, \beta)$ can be motivated as follows. Let Z_i be the number of reads covering the first occurrence of $\mathbf{w}$ in the i-th clump. We can reasonably assume that the read will cover the whole clump as the clump size is generally not long. Then we may approximate

$$\mathcal{N}_{\mathbf{w}}(M, n, \beta) \approx \sum_{i=1}^{N_{C,n}} X_i Z_i. \tag{5}$$

We note that the above equation may slightly over-estimate the number of occurrences of $\mathbf{w}$ in the M reads since we only require that the reads cover the first $\mathbf{w}$, not the whole clump. However, the approximation is reasonable since the sequence reads are generally much longer than the length of clumps and the reads covering the first $\mathbf{w}$ are most likely covering the whole clump.

Next, we study the distribution of Z_i. If the i-th clump starts at j, then the number of reads containing the first $\mathbf{w}$ in the clump is a binomial random variable $B(M, \lambda_{j-\beta+w} + \cdots + \lambda_j)$ which is asymptotically Poisson with mean $\Lambda_j = M \sum_{i=(j-\beta+w)\vee 1}^{j} \lambda_i$. Since the occurrences of clumps follow asymptotically a Poisson process, the starting point of the i^{th} clump is approximately uniformly distributed in $[1, n - w + 1]$. Thus, the independent random variables $\tilde{Z}_i$ with distribution

$$\mathbb{P}\{\tilde{Z}_i = k\} = \frac{1}{n-w+1} \sum_{j=1}^{n-w+1} \frac{(\Lambda_j)^k}{k!} \exp(-\Lambda_j) \tag{6}$$

are a reasonable approximation to the random variables Z_i.

The next theorem makes the heuristic argument precise. Recall that the *total variation distance* between two $\mathbb{Z}_+$-valued random variables X and Y (defined on the same probability space) is defined by

$$d_{\mathrm{TV}}(X, Y) = \sup_{A \in \mathbb{Z}_+} |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)|.$$

Thus, if the total variation distance between X and Y is small, then for any subset, A, of the nonnegative integers, the difference between the probability for X to be in A and that for Y is also small. A bound in total variation distance can be applied to get conservative p-values for counts via the formula

$$\mathbb{P}(X \le x) \le \mathbb{P}(Y \le x) + d_{\mathrm{TV}}(X, Y).$$

To state the results we need some more notation. Let $\alpha = \alpha_2$ be the second-largest eigenvalue of the transition matrix T; the Perron-Frobenius Theorem ensures that $|\alpha| < 1$. Let D be the matrix with the eigenvalues of T on the diagonal, ordered such that the first entry is $\alpha_1 = 1$, and zero entries everywhere else. Then we decompose $T = PDP^{-1}$ such that the first column of P is $(1, 1, \ldots, 1)^T$. For all $t \in \{0, 1, \ldots, L-1\}$, let J_t denote the $L \times L$ matrix such that all its entries are equal to 0 except $J_t(t, t) = 1$, and we define

$$Q_t := PJ_tP^{-1}.$$

Let $\pi(x)$ be the probability of letter x and

$$\gamma_2(v) = \sum_{x, y \in \{0, \ldots, L-1\}} \pi(x) \max_{a, b \in \{0, \ldots, L-1\}} \left(\frac{1}{\pi(b)} \sum_{(t, t') \neq (1, 1)} \left| \frac{\alpha_t^v \alpha_{t'}^v}{\alpha^v} Q_t(x, b) Q_{t'}(a, y) \right| + \sum_{t=2}^{|L|} \left| \frac{\alpha_t^{5v-3}}{\alpha^v} Q_t(x, y) \right| \right).$$

Let $\pi_{\min}$ be the smallest value of $\{\pi(a), a \in \{0, 1, \cdots, L-1\}\}$. Put

$$\begin{aligned} B(T, \mathbf{w}, n) = {} & (n-w+1)\widehat{\mu}(\mathbf{w}) \left(2(w-1)P(\mathbf{w}) + (6w-5)\hat{\mu}(\mathbf{w}) + \gamma_2(w)|\alpha|^w \right) \\ & + 2(n-w+1) \left\{ \frac{P^2(\mathbf{w})}{\pi(\mathbf{w}_1)} \sum_{s=1}^{2w-2} T^s(\mathbf{w}_w, \mathbf{w}_1) + \frac{\hat{\mu}(\mathbf{w})}{\mu_{\min}} ((w-2)P(\mathbf{w}) + \hat{\mu}(\mathbf{w})) \right\}. \end{aligned} \tag{7}$$

Theorem 2.2 *Let $\widetilde{NC}_n$ have Poisson distribution with mean $\Lambda(\mathbf{w})$, let $\tilde{Z}_i, i=1, \ldots, n-w+1$ have distribution* (6)*, let $X_i, i=1, \ldots, n-w+1$ have distribution* (3) *and assume that all these variables are independent. Then*

$$d_{\mathrm{TV}}\left(\mathcal{N}_{\mathbf{w}}(M, n, \beta), \sum_{i=1}^{\widetilde{NC}_n} X_i \tilde{Z}_i\right) \leq B(T, \mathbf{w}, n) + 2M \sum_{i=1}^{\bar{n}} \lambda_i^2 + 2(w-1)(P(\mathbf{w}) - \hat{\mu}(\mathbf{w})).$$

Here $B(T, \mathbf{w}, n)$ is given in (7).

Let $\Lambda_j = M \sum_{i=(j-\beta+w)\vee 1}^{j} \lambda_i$. Then the probability $g_k = \mathbb{P}(\mathcal{N}_{\mathbf{w}}(M, n, \beta) = k)$ can be calculated using the recursion (Panjer, 1981, Willmot and Panjer, 1987)

$$g_k = \frac{\Lambda(\mathbf{w})}{k} \sum_{j=1}^{k} j f_j g_{k-j} \tag{8}$$

with initial value $g_0 = \exp((f_0 - 1)\Lambda(\mathbf{w}))$, $f_j = \mathbb{P}(X_i \tilde{Z}_i = j) = \sum_{l \cdot m = j} \mathbb{P}(X_i = l)\mathbb{P}(\tilde{Z}_i = m)$, and $f_0 = \frac{1}{n-w+1} \sum_{i=1}^{n-w+1} \exp(-\Lambda_i)$.

While $B(T, \mathbf{w}, n)$ has a complicated expression, when the Markov chain is reasonably well mixed then its leading term will be of the order $\eta \hat{\mu}(\mathbf{w}) w P(\mathbf{w})$. The bound in Theorem 2.2 will be small when, firstly, the compound Poisson approximation for intervals of length β is good; secondly, the distribution of starting points of reads is relatively homogeneous; thirdly, the number M of reads is not too large compared to n; and fourthly, $\sum_{p \in \mathcal{P}'(\mathbf{w})} P(\mathbf{w}^{(p)}\mathbf{w})$ is small.

Theorem 6.6.4 in Reinert et al. (2005) gives the analogous approximation for counts of different words $\mathbf{w}_1, \ldots, \mathbf{w}_r$, where r is an integer. The bounds are of similar flavor but involve more notation which considers the possible overlaps between different words. We omit the result here.

2.4. Extending the approximations to the double-strand model

In the above subsections, we assume that the sequence under study is single-stranded for simplicity. However, DNA sequences are double-stranded and the sequence reads from NGS can come from either strand and it is not known which strand the reads come from. To take both strands into consideration, we consider both the reads and their complements. Among the M pairs of reads, the number of occurrences of $\mathbf{w}$, $\tilde{\mathcal{N}}_{\mathbf{w}}(M, n, \beta)$, is equal to the number of occurrences of the complement of $\mathbf{w}$, $\tilde{\mathcal{N}}_{\bar{\mathbf{w}}}(M, n, \beta)$, because we consider the complement of each read. Next we study the distribution of $\tilde{\mathcal{N}}_{\mathbf{w}}(M, n, \beta)$ for any word pattern $\mathbf{w}$ by considering the following scenarios.

We first assume a palindrome such that $\mathbf{w} = \bar{\mathbf{w}}$, for example, $\mathbf{w}$ = ACGT or CGCG. For such word patterns, it is obvious that $\tilde{\mathcal{N}}_{\mathbf{w}}(M, n, \beta) = 2\mathcal{N}_{\mathbf{w}}(M, n, \beta)$. Next we assume $\mathbf{w} \neq \bar{\mathbf{w}}$. For each pair of complementary reads, we consider the one from the forward strand. Thus, we have a new set of M reads all from the forward strand. Note that the word pattern $\mathbf{w}$ occurs in one of the strands of a pair of complementary reads if the forward read contains either $\mathbf{w}$ or $\bar{\mathbf{w}}$. Thus, the total number of occurrences of $\mathbf{w}$ in the M pairs of complementary reads equals the number of occurrences of either $\mathbf{w}$ or $\bar{\mathbf{w}}$ along the forward reads. Thus, we are interested in the joint word counts of $\mathbf{w}$ and its complement $\bar{\mathbf{w}}$, but in contrast to Reinert and Schbath (1998) we allow for mixed clumps of occurrences, that is, the clumps can be composed of combinations of $\mathbf{w}$ and its complement. A compound Poisson approximation, with bounds, for the joint count of $\mathbf{w}$ and its complement can be found in Roquain and Schbath (2007). The approximation is valid only for non-palindromes; it also requires that the word and its complement have non-zero probability of appearing in the sequence. Here we illustrate how such a compound Poisson approximation for word counts in single reads can be combined for NGS data.

Let

$$C_k = \left\{ \hat{\mathbf{w}}_1^{(p_1)} \hat{\mathbf{w}}_2^{(p_2)} \cdots \hat{\mathbf{w}}_{k-1}^{(p_{k-1})} \hat{\mathbf{w}}, p_i \in \mathcal{P}'(\hat{\mathbf{w}}_i, \hat{\mathbf{w}}_{i+1}) \right\},$$

where $\hat{\mathbf{w}}_i, i = 1, 2, \cdots$ can be either $\mathbf{w}$ or $\bar{\mathbf{w}}$, and

$$\mathcal{P}(\mathbf{u}, \mathbf{v}) = \{p : \mathbf{u}_{p+1} = \mathbf{v}_1, \mathbf{u}_{p+2} = \mathbf{v}_2, \cdots, \mathbf{u}_u = \mathbf{v}_{u-p}\},$$

and $\mathcal{P}'(\mathbf{u}, \mathbf{v})$ is a subset of $\mathcal{P}(\mathbf{u}, \mathbf{v})$ by removing those that are multiple of other numbers. By the definition, we have, for any word pattern $\mathbf{w}$,

$$\mathcal{P}(\mathbf{w}, \mathbf{w}) = \mathcal{P}(\bar{\mathbf{w}}, \bar{\mathbf{w}}).$$

Let $C_k(\mathbf{w})$ be the subset of C_k such that the first word $\hat{\mathbf{w}}_1 = \mathbf{w}$, and $C_k(\bar{\mathbf{w}})$ be the subset of C_k such that the first word $\hat{\mathbf{w}}_1 = \bar{\mathbf{w}}$. Define $\mathcal{S}_{\mathbf{w}} = \{\mathbf{w}^{(p)}, p \in \mathcal{P}'(\mathbf{w}, \mathbf{w})\} \bigcup \{\bar{\mathbf{w}}^{(p)}, p \in \mathcal{P}'(\bar{\mathbf{w}}, \mathbf{w})\}$ and $\mathcal{S}_{\bar{\mathbf{w}}} = \{\mathbf{w}^{(p)}, p \in \mathcal{P}'(\mathbf{w}, \bar{\mathbf{w}})\} \bigcup \{\bar{\mathbf{w}}^{(p)}, p \in \mathcal{P}'(\bar{\mathbf{w}}, \bar{\mathbf{w}})\}$.

Given the above notation, we consider the distribution of clumps starting with $\mathbf{w}$ and $\bar{\mathbf{w}}$, respectively. Here a clump is defined as a maximum region with overlapping occurrences of either $\mathbf{w}$ or $\bar{\mathbf{w}}$. Note that a clump starting with $\mathbf{w}$ occurs at a position i if 1) $\mathbf{w}$ occurs at position i, 2) sequences in $\mathcal{S}_{\mathbf{w}}$ do not occur just before i. Thus, a $\mathbf{w}$-clump starts at a particular position with probability

$$\hat{\mu}'(\mathbf{w}) = P(\mathbf{w}) - P(\mathcal{S}_{\mathbf{w}}\mathbf{w}). \tag{9}$$

Similarly, the probability that a $\bar{\mathbf{w}}$-clump starts at a particular position

$$\hat{\mu}'(\bar{\mathbf{w}}) = P(\bar{\mathbf{w}}) - P(\mathcal{S}_{\bar{\mathbf{w}}}\bar{\mathbf{w}}). \tag{10}$$

We refer to the clumps starting with $\mathbf{w}$ the $\mathbf{w}$-clumps and those starting with $\bar{\mathbf{w}}$ the $\bar{\mathbf{w}}$-clumps. Both the $\mathbf{w}$-clumps and $\bar{\mathbf{w}}$-clumps can be approximated by a Poisson process with parameters $\hat{\mu}'(\mathbf{w})$ and $\hat{\mu}'(\bar{\mathbf{w}})$, respectively. The joint of the two approximate Poisson processes can again be approximated by a Poisson process with parameters $\hat{\mu}'(\mathbf{w}) + \hat{\mu}'(\bar{\mathbf{w}})$. Thus, the number of clumps including both the $\mathbf{w}$-clumps and the $\bar{\mathbf{w}}$-clumps, $N'_{C,n}$, can be approximated by a Poisson random variable with mean $(n - w + 1)(\hat{\mu}'(\mathbf{w}) + \hat{\mu}'(\bar{\mathbf{w}}))$.

We order the $\mathbf{w}$- and $\bar{\mathbf{w}}$-clumps from the 5′-end to the 3′-end. Let $I_i = 1$ and $I_i = 0$ be the events that the i-th clump is a $\mathbf{w}$-clump and $\bar{\mathbf{w}}$-clump, respectively. We have

$$\mathbb{P}(I_i = 1) = 1 - \mathbb{P}(I_i = 0) = \frac{\hat{\mu}'(\mathbf{w})}{\hat{\mu}'(\mathbf{w}) + \hat{\mu}'(\bar{\mathbf{w}})}. \tag{11}$$

Next we study the distribution of the number of occurrences of $\mathbf{w}$ or $\bar{\mathbf{w}}$ in $\mathbf{w}$-clumps and $\bar{\mathbf{w}}$-clumps, separately. A k-clump starting with $\mathbf{w}$ is referred as a k-$\mathbf{w}$-clump. Similarly, a k-clump starting with $\bar{\mathbf{w}}$ is referred as a k-$\bar{\mathbf{w}}$-clump. Then a k-$\mathbf{w}$-clump occurs at a specific position i if (1) $C_k(\mathbf{w})$ occurs at position i, (2) sequences in $\mathcal{S}_{\mathbf{w}}$ do not occur just before i, and (3) $C_{k+1}(\mathbf{w})$ does not occur at position i. Note that when we deduct the probability of the second and the third events, we deduct the probability of the event $SC_{k+1}(\mathbf{w})$ twice. Thus, we need to add the probability of this event, giving

$$\hat{\mu}'_k(\mathbf{w}) = \mathbb{P}(k\text{-}\mathbf{w}\text{-clump at a position}) = P(C_k(\mathbf{w})) - P(\mathcal{S}_{\mathbf{w}}C_k(\mathbf{w})) - P(C_{k+1}(\mathbf{w})) + P(\mathcal{S}_{\mathbf{w}}C_{k+1}(\mathbf{w})). \tag{12}$$

Similarly, we have for $\bar{\mathbf{w}}$

$$\hat{\mu}'_k(\bar{\mathbf{w}}) = \mathbb{P}(k\text{-}\bar{\mathbf{w}}\text{-clump at a position}) = P(C_k(\bar{\mathbf{w}})) - P(\mathcal{S}_{\bar{\mathbf{w}}}C_k(\bar{\mathbf{w}})) - P(C_{k+1}(\bar{\mathbf{w}})) + P(\mathcal{S}_{\bar{\mathbf{w}}}C_{k+1}(\bar{\mathbf{w}})). \tag{13}$$

Let X_i and $\bar{X}_i$ be the numbers of occurrences of $\mathbf{w}$ or $\bar{\mathbf{w}}$ in a $\mathbf{w}$-clump and $\bar{\mathbf{w}}$-clump, respectively. Then the distributions of X_i and $\bar{X}_i$ are

$$\mathbb{P}(X_i = k) = \frac{\hat{\mu}'_k(\mathbf{w})}{\hat{\mu}'(\mathbf{w})}, \quad \mathbb{P}(\tilde{X}_i = k) = \frac{\hat{\mu}'_k(\bar{\mathbf{w}})}{\hat{\mu}'(\bar{\mathbf{w}})}. \tag{14}$$

Similar as above, let $\widetilde{Z}_i$ be the number of reads covering the first occurrence of $\mathbf{w}$ or $\bar{\mathbf{w}}$ in the i-th clump. The distribution of $U_i = (I_i X_i + 1(1 - I_i)\bar{X}_i)\widetilde{Z}_i$ can be calculated by

$$f'_j = \mathbb{P}(U_i = j) = \sum_{l,m;l\times m=j} \mathbb{P}(\widetilde{Z}_i = m)\left(\frac{\hat{\mu}'(\mathbf{w})}{\hat{\mu}'(\mathbf{w}) + \hat{\mu}'(\bar{\mathbf{w}})}\mathbb{P}(X_i = l) + \frac{\hat{\mu}'(\bar{\mathbf{w}})}{\hat{\mu}'(\mathbf{w}) + \hat{\mu}'(\bar{\mathbf{w}})}\mathbb{P}(\bar{X}_i = l)\right).$$

The number of occurrences of $\mathbf{w}$ along the M pairs of reads can be approximated by

$$\widetilde{\mathcal{N}}_{\mathbf{w}}(M, n, \beta) = \sum_{i=1}^{N'_{C,n}} U_i = \sum_{i=1}^{N'_{C,n}} (I_i X_i + (1 - I_i)\bar{X}_i)\widetilde{Z}_i. \tag{15}$$

The argument is made precise in the next proposition.

Theorem 2.3. *Let* $\widetilde{NC}'_n$ *have a Poisson distribution with mean* $(n-w+1)(\hat{\mu}'(\mathbf{w})+\hat{\mu}'(\mathbf{w}))$ *let* $\tilde{Z}_i, i=1, \ldots, n-w+1$ *have distribution* (6)*, let* $X_i, i=1, \ldots, n-w+1$ *have distribution* (14)*, let* $I_i, i=1, \ldots, n-w+1$ *have distribution* (11) *and assume that all these variables are independent. Then*

$$d_{\mathrm{TV}}\left(\mathcal{N}_{\mathbf{w}}(M, n, \beta), \sum_{i=1}^{\widetilde{NC}'_n}(I_iX_i+(1-I_i)\bar{X}_i)\widetilde{Z}_i\right) \leq C(T, \mathbf{w}, n)+2M\sum_{i=1}^{\bar{n}}\lambda_i^2+2w(P(\mathbf{w})+P(\bar{\mathbf{w}})).$$

Here

$$C(T, \mathbf{w}, n) := CnwP(\mathbf{w})^2+C'n(P(\mathbf{w})+P(\bar{\mathbf{w}}))|\alpha|^w,$$

where $C > 0$ *and* $C' > 0$ *are two explicit constants that only depend on the transition matrix T, and* α *is the second largest eigenvalue in modulus of T.*

2.5. The approximate power of detecting enriched patterns using the compound Poisson approximation for the distribution of the number of occurrences of word patterns

The framework for the normal and compound Poisson approximations for the number of occurrences of word patterns can equally be applicable to sequences generated by hidden Markov models. In particular, a regulatory sequence with many instances of transcription factor binding sites can be modeled by a hidden Markov model as in Zhai et al. (2010). Specifically, the long background sequence is modeled as an i.i.d. sequence. In addition, instances of a motif with a given position weight matrix are randomly inserted into the background sequence with probability $1-\rho$ at each position. We refer to $1-\rho$ as motif density. Next generation sequencing is then used to sample M reads from the long sequence as modeled in Subsection 2.1. Based on the reads, we want to test the null hypothesis $H_0: \rho = 1$, *i.e*, no motif instances are inserted, versus the alternative hypothesis $H_1: \rho < 1$, *i.e*, motif instances are inserted in the underlying background sequence. Consider the dominant pattern $\mathbf{w}^{(d)}$ in the motif consisting of the nucleotide with the highest probability in each position. We can use $\mathcal{N}_{\mathbf{w}(d)}(M, n, \beta)$ as a statistic to test the hypotheses. For a given type I error α, we can obtain a threshold t_α, that is, the smallest value of t such that

$$\mathbb{P}_1(\mathcal{N}_{\mathbf{w}(d)}(M, n, \beta) \geq t) \leq \alpha, \tag{16}$$

based on the theory for $\mathbb{P}_1$ developed above, where $\mathbb{P}_1$ is the approximate probability distribution of $\mathcal{N}_{\mathbf{w}(d)}(M, n, \beta)$ when no motifs are inserted, *i.e*. $\rho = 1$.

The approximate power of the test statistic when $\rho < 1$ is given by

$$\text{power} = \mathbb{P}_\rho(\mathcal{N}_{\mathbf{w}(d)}(M, n, \beta) \geq t_\alpha), \tag{17}$$

where $\mathbb{P}_\rho$ is the probability distribution under the alternative model. The approximate power for detecting the enrichment of certain patterns under the double-strand model can be calculated similarly. In the following, for convenience we use the term ''power'' to mean approximate power.

3. RESULTS: SIMULATION STUDIES AND BIOLOGICAL APPLICATIONS

3.1. Evaluation of the accuracy of the normal and compound Poisson approximations using simulations

We study nucleotide sequences consisting of four states (A, C, G, T) and consider three relatively short patterns (''TAT,'' ''ACGT,'' and ''CGCG'') and two relatively long patterns (''ACGTATC'' and ''AAGAAGAA''). The pattern ''ACGT'' does not have any periods and the patterns ''TAT'' and ''CGCG'' have a period 2. The pattern ''ACGTATC'' does not have any periods and the pattern ''AAGAAGAA'' has three periods {3, 6, 7} with principal periods {3, 7}. For each pattern, we compare the histogram of the simulated number of occurrences of a pattern with the theoretical compound Poisson approximation probability mass function given in Section 2. For patterns having relatively high number of occurrences, e.g, mean at least 50 in the cases we consider, we also plot the density function of the normal approximation.

In all our simulations, we use the following parameters: the nucleotide frequencies of (A, C, G, T) are (a) GC-rich (0.15, 0.35, 0.35, 0.15), (b) uniform (0.25, 0.25, 0.25, 0.25), and (c) GC-poor (0.35, 0.15, 0.15,

0.35). These settings allow us to see the effect of nucleotide frequency on the accuracy of the theoretical approximations. The sequence length n is chosen as either 5000 for the two short patterns or 20,000 for the two long patterns. The number of sequence reads M is either 500 for the two short patterns or 4000 for the two long patterns. The read length $\beta = 80$. We assume that the sequence reads are either homogeneously or heterogeneously chosen from the long sequence. In heterogeneous sampling of the reads, we divide the sequence into 100 consecutive blocks. For each block, we sample a random number from the gamma distribution $\Gamma(1, 20)$ and the sampling probability λ_i for each position in the block is proportional to the chosen number (Zhang et al., 2008). We consider both the single- and double-strand models in our simulation studies. The number of simulations for each case is 10,000. Note that the sequence length and the number of reads simulated here are much smaller than the corresponding values in real studies. These numbers are chosen to save computational time. The qualitative results should hold for much longer sequences and higher number of reads.

Due to page limitations, we present the figures for the simulation results as Figures S1–S9 in the Supplementary Material (see www.liebertonline.com/cmb for Supplementary Material). We make the following observations. First, when the average number of occurrences of the pattern of interest is relatively large, for example, greater than 500, both the normal and compound Poisson approximations work well. However, for most of the cases we considered in this study, the compound Poisson approximation outperforms the normal approximation. Second, when the sequence reads are heterogeneously sampled from the long background sequence, the range of the number of occurrences of the patterns will be larger than that under the homogeneous sampling scheme. Third, the distribution of the number of occurrences of patterns can have multiple peaks under some situations and the compound Poisson approximation can accurately capture such features.

3.2. Power studies using simulations

We next study the power of detecting enriched patterns when such patterns indeed occur more frequently than expected. For this we consider sequences with random instances of a motif on a background i.i.d. sequences as described in Subsection 2.5. Under this model, a motif instance is inserted at each position of the background sequence with probability $1 - \rho$ (motif density). Thus, the consensus pattern of the motif is enriched compared to the background. The background sequence models and the inserted motifs ''TAT,'' ''ACGT,'' ''CGCG,'' ''ACGTATC,'' and ''AAGAAGAA'' are all the same as in Subsection 3.1. For a type I error $\alpha = 0.025$, we first use Equation 16 to find the threshold t_α based on the approximate distribution for $\mathcal{N}_{\mathbf{w}(d)}(M, n, \beta)$ under the null model $\rho = 1$. The theoretical power under the alternative model that $\rho < 1$ is calculated using Equation 17. We run 10,000 simulations based on the alternative model and record the number of occurrences of the corresponding pattern. The simulated power is approximated by the fraction of times that $\mathcal{N}_{\mathbf{w}(d)}(M, n, \beta) \geq t_\alpha$, where $\mathbf{w}^{(d)}$ is the consensus pattern of the motif and it is the inserted pattern in our simulations.

Tables 1 and 2 compare the theoretical and the simulated power of $\mathcal{N}_{\mathbf{w}}(M, n, \beta)$ for detecting the corresponding enriched patterns for different values of motif density $1 - \rho$ for the patterns: ''TAT,'' ''ACGTATC,'' and ''AAGAAGAA'' under the single- and double-strand GC-poor models, respectively. The power of detecting the patterns ''ACGT'' and ''CGCG'' using the single- and double-strand model is the same because the counts for the double-strand model is twice the count for the single-strand model. The power results for these two patterns under the GC-poor model are given as Table S1 and the complete results for the GC-rich and uniform background models are given as Tables S2–S7 in Supplementary Material. The following conclusions can be obtained from the tables. First, the threshold value calculated from Equation 16 is conservative in that the simulated type I error rate is smaller than the specified type I error α in most of the situations. Second, the theoretical power given in Equation 17 is very close to the simulated power when the theoretical power is relatively large (e.g, greater than 50%). Third, the power of detecting enriched patterns under heterogeneous read sampling is smaller than that under homogeneous read sampling.

3.3. Applications to a ChIP-Seq data set in Valouev et al. (2008)

Now we apply the theory to a ChIP-Seq data set using transcription factor GABP in Valouev et al. (2008). We consider the promoter region of Nuclear Matrix Transcription Factor 4 gene (ZNF384) between position 6,667,900 and position 6,669,500 (a total of 1600 bp) on human chromosome 12, NCBI build 36.

Table 1. Comparison of the Simulated and Theoretical Power (×100) for Patterns TAT, ACGTATC, and AAGAAGAA Under the Single-Strand GC-Poor Model

			Scaled motif density									
		Threshold	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Homogeneous read sampling												
TAT	Simulation	1966	1.2	2.2	4.7	8.9	14.5	21.9	33.3	45.1	55.2	67.1
	Theory	1966	2.4	4.8	8.4	13.7	20.7	29.3	39.1	49.4	59.5	68.9
ACGTATC	Simulation	54	1.8	14.8	32.0	51.8	70.8	83.5	89.3	93.5	97.1	98.4
	Theory	54	2.4	14.9	34.9	55.3	71.8	83.3	90.6	94.9	97.4	98.7
AAGAAGAA	Simulation	49	2.2	15.6	38.8	56.3	74.2	87.0	92.3	95.3	97.0	99.2
	Theory	49	2.3	16.6	38.2	58.9	74.9	85.6	92.1	95.9	97.9	99.0
Heterogeneous read sampling												
TAT	Simulation	2026	1.3	2.1	5.1	7.9	11.2	16.5	24.2	32.3	40.3	49.6
	Theory	2026	2.4	4.2	6.7	10.1	14.6	20.1	26.5	33.8	41.6	49.6
ACGTATC	Simulation	70	2.4	9.7	19.6	33.6	46.0	62.6	72.7	81.3	87.9	93.3
	Theory	70	2.4	9.7	21.4	35.4	49.5	62.3	72.9	81.2	87.4	91.7
AAGAAGAA	Simulation	63	2.0	12.2	27.5	38.7	51.8	67.7	76.0	85.3	89.8	93.0
	Theory	63	2.4	11.2	24.4	39.5	54.1	66.7	76.8	84.4	89.8	93.5

The sequence length $n = 5000$, the number of reads $M = 500$, and the scaled motif density = $1000(1 - \rho)$ for the pattern TAT. For the two long patterns, the sequence length $n = 20000$, the number of reads $M = 4000$ and the scaled motif density = $20000(1 - \rho)$. The read length $\beta = 80$. Type I error $\alpha = 2.5\%$. The ''Threshold'' is obtained using Equation 16 based on the theoretical approximate distribution of the number of occurrences of a pattern under the null model. The number of simulations is 10,000.

The gene ZNF384 has been shown to be the regulatory target of GABP and the region is enriched with ChIP-Seq reads as shown in the supplementary materials of Valouev et al. (2008). Our objective is to show the applicability of the theory developed in this article, not as a comparison with other computational methods of peak calling for ChIP-Seq data. The position weight matrix (PWM) of the GABP binding side is given in Table S8 (JASPAR, http://jaspar.cgb.ki.se/; ID number MA0062.2) in Supplementary Material. The consensus sequence formed by the dominant nucleotide at each position is ''CCGGAAGTGGC''.

In a typical ChIP-Seq experiment, DNA regions of interest are sheared into short fragments and the specific DNA fragments interacting with the protein of interest are isolated by immuno-precipitation. Then NGS is used to sequence either end of the sequence. These end sequences are referred as tag sequences. In

Table 2. Comparison of the Simulated and Theoretical Power (×100) for Patterns TAT, ACGTATC, and AAGAAGAA Under the Double-Strand GC-Poor Model

			Scaled motif density									
		Threshold	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Homogeneous read sampling												
TAT	Simulation	3875	0.9	1.5	2.9	5.7	8.8	13.9	21.5	30.1	37.9	48.2
	Theory	3875	2.49	4.3	6.9	10.6	15.4	21.4	28.3	36.0	44.2	52.4
ACGTATC	Simulation	82	2.3	13.8	37.1	53.8	73.6	83.5	91.8	94.7	98.1	98.9
	Theory	82	2.4	16.1	37.4	58.3	74.5	85.4	92.0	95.8	97.9	99.0
AAGAGAA	Simulation	73	1.9	10.5	26.3	43.3	60.7	72.9	83.5	89.3	93.3	97.0
	Theory	73	2.3	11.2	26.4	44.2	60.9	74.3	84.0	90.4	94.5	97.0
Heterogeneous read sampling												
TAT	Simulation	3982	1.2	1.9	2.9	4.5	6.9	10.6	16.8	21.5	27.5	33.4
	Theory	3982	2.49	3.8	5.7	8.1	11.2	14.9	19.4	24.4	30.0	36.0
ACGTATC	Simulation	101	2.2	7.7	12.9	27.7	37.5	51.3	60.0	68.9	79.2	83.7
	Theory	101	2.46	7.4	15.5	26.0	37.8	49.7	60.7	70.4	78.3	84.5
AAGAAGAA	Simulation	91	2.48	7.1	15.6	26.1	38.3	53.5	66.2	71.9	79.1	86.2
	Theory	91	2.4	7.9	16.9	28.2	40.6	52.8	63.8	73.2	80.7	86.4

The parameters are the same as in Table 1.

Valouev et al. (2008), the tag sequences are of length 25 bp. Since the tag sequences from ChIP-Seq can come from either the forward or the reverse strand of the selected fragments and the tag sequences may not contain the GABP binding sites, we extend both the forward and reverse strands to the whole sequence fragments as follows. It was estimated in Valouev et al. (2008) that the median read length in the GABP data set is 56 bp with mean around 57 bp. So we extend the forward strand by 31 bp in the forward direction. We also extend the reverse strand in the reverse direction by 31 bp so that each tag is associated with a read of 56 bp.

We analyze three different read data sets mapped to the forward strand only, the reverse strand only, and both the forward and the reverse strands combined. For each k-tuple (k = 6), we first approximate the p-value corresponding to the k-tuple using the control (Rx-noIP) data and the compound Poisson approximation. In such calculations, we use the nucleotide frequencies calculated from the extended reads. The distribution of the reads along the genomic region is estimated empirically by the fraction of reads starting from each position as follows. Since the number of reads starting from individual positions is generally small and the estimated distribution of reads λ_i using the number of reads starting at the i-th position is not reliable, we estimate λ_i within a window using the following approach. For a given window size S, we estimate λ_i by the average number of reads starting at the positions within the window of size S centered at i, in other words,

$$\hat{\lambda}_i = \frac{\sum_{i'=i-[S/2]}^{i+[S/2]} M_i}{S \times M},$$

where M_i is the number of reads starting at position i and M is the total number of reads. Using these estimated parameters, we can approximate the p-value corresponding to each k-tuple.

We use our approach to analyze both the control and the ChIP-Seq data. We expect that no k-tuples are significant for the control data while the dominant patterns in the motif should be enriched in the ChIP-Seq data set. Different window sizes $S = 1$ to 50 by step 5 are used and the results are similar. Table 3 presents the top 10 k-tuples using $k = 6$ with the smallest p-values when $S = 20$ using the reads mapped to the forward strand, reverse strand, and both strands for the control data, and Table 4 presents the results based on the ChIP-Seq data.

For family-wise type I error 0.05, using the Bonferroni correction, only 6-tuples with p-value less than $0.05/4^6 \approx 1.25 \times 10^{-5}$ are declared as significant. With this criterion, one 6-tuple ''CACTTC'' was identified as significant using the reads mapped to the forward strand based on the control data. The tuple ''CACTTC'' is complementary to the dominant pattern at positions [4,9]. The next pattern with relatively small p-value, although not significant using our criterion, is ''ACTTCC'' which is complementary to the dominant pattern at positions [5,10]. From the two patterns, it is possible to construct a consensus sequence of seven nucleotides ''CACTTCC'' which is complementary to the dominant pattern at positions [4,10]. We see some GABP motif signals in the control data set.

TABLE 3. TOP 10 k-TUPLES ($k = 6$) WITH SMALLEST p-VALUES USING READS MAPPED TO THE FORWARD, REVERSE, AND BOTH FORWARD AND REVERSE STRANDS FOR THE CONTROL DATA SET WITH TRANSCRIPTION FACTOR GABP

Forward		*Reverse*		*Combined*		
6-tuple	p-*value*	*6-tuple*	p-*value*	*6-tuple*	*Complement*	p-*value*
CACTTC	5.82E-06	GAAGTG	9.53E-05	CACTTC	GAAGTG	6.01E-07
ACTTCC	0.000349	GTGAGT	0.000584	CTATAG	CTATAG	5.76E-05
CTTCTG	0.000956	AGTCCT	0.000678	ACTTCC	GGAAGT	6.98E-05
TCCTTG	0.000956	GGAGGG	0.000731	CCCTCC	GGGAGG	1.11E-04
CCACTT	0.001367	CCGGAA	0.000961	CCGGAA	TTCCGG	1.56E-04
CCTTCC	0.001672	GTCCTC	0.001143	CAGAAG	CTTCTG	2.24E-04
CCTTGC	0.001743	AAGTGG	0.001221	ACTTCT	AGAAGT	3.26E-04
TTCCGG	0.001816	GAAGAA	0.001353	GCTATA	TATAGC	6.02E-04
CTCCTT	0.001828	CTATAG	0.001355	CGGAAG	CTTCCG	8.47E-04
CCTTGT	0.001899	GCTATA	0.001355	AAGTGG	CCACTT	1.02E-03

Table 4. Top 10 k-Tuples ($k = 6$) with Smallest p-Values Using Reads Mapped to the Forward, Reverse, and Both Forward and Reverse Strands for the ChIP-Seq Data Set with Transcription Factor GABP

Forward		*Reverse*		*Combined*		
6-tuple	p-*value*	*6-tuple*	p-*value*	*6-tuple*	*Complement*	p-*value*
CACTTC	5.55E-11	CGGAAG	2.06E-08	ACTTCC	GGAAGT	1.78E-15
ACTTCC	1.75E-10	CCGGAA	2.1E-08	CAGAAG	CTTCTG	1.78E-15
TTCCGG	4.97E-10	CTTCCG	2.75E-07	CGGAAG	CTTCCG	2.33E-15
CTTCCG	6.47E-10	TAGCGG	3.26E-07	CACTTC	GAAGTG	3.55E-15
CAGAAG	8.82E-07	GAAGCT	5.42E-07	GCAGAA	TTCTGC	7.44E-15
GCAGAA	1.16E-06	CTAGCG	1.2E-06	CCGGAA	TTCCGG	1.42E-14
AAATAG	1.25E-05	CACTTC	1.31E-06	CTATAG	CTATAG	2.80E-13
GAAATA	1.3E-05	ACTTCC	1.48E-06	AAGTGA	TCACTT	4.21E-12
TCACTT	1.32E-05	GAAGTG	1.53E-06	GCGGAA	TTCCGC	6.54E-12
ACACTT	2.45E-05	GGAAGT	1.77E-06	CCGCTA	TAGCGG	1.46E-10

For the ChIP-Seq data, six 6-tuples are significant based on reads mapped to the forward strand. The top four 6-tuples with p-value at most 6.47×10^{-10}—"CACTTC," "ACTTCC," "TTCCGG," and "CTTCCG"—are complementary to the dominant patterns at positions [4,9], [3,8], [1,6], and [2,7], respectively. From these four patterns, we are able to construct a consensus sequence of 9 nucleotides "CACTTCCGG." Similar observations can be made based on reads mapped to the reverse strand and both strands. We also carried out similar studies using $k = 4$ and 5, and the results are similar (data not shown).

4. DISCUSSION

In this article, we study the distribution of the number of occurrences of patterns in sequence reads randomly sampled from long Markovian sequences. This problem comes naturally from the analysis of sequence reads generated from NGS including ChIP-Seq and RNA-Seq. In this article, we first develop probabilistic models for the background sequences and the sampling of sequence reads using NGS. The background sequence is modeled as a Markovian sequence. Each sequence read starts from the i-th position of the background sequence with probability λ_i and the sampling of sequence reads from the background sequence is assumed to be independent. Based on the model, we study the limit distribution of the number of occurrences of any k-tuple patterns. Two approximate distributions are considered. We assume throughout the paper that both the background sequence length and the number of sequence reads are large and that the sequence reads do not concentrate on particular regions of the background sequence. We first give a normal approximation for the number of occurrences of frequent patterns and provide formulas to calculate the mean and variance of the approximate normal distribution. For relatively rare patterns, we provide a new compound Poisson approximation for the number of occurrences. Simulation studies are first used to evaluate the theoretical results, and it is shown that the compound Poisson approximation seems to work well in most of the situations. The compound Poisson approximation is then used to analyze ChIP-Seq data mapped to the promoter region of the gene ZNF384 using transcription factor GABP. Surprisingly, we found GABP binding motif signals in the control data set indicating some ChIP residue effect even within the control data. With the ChIP-Seq data, we are able to recover the consensus patterns of the motif.

Despite the usefulness of the models and the approximations, there are some limitations. First, we assume that the background sequence follows a homogeneous Markov chain. In reality, the sequence to be sequenced may be heterogeneous with different regions following varied Markov models. If we have some idea about the composition of the nucleotides at different parts of the sequence, hidden Markov models can potentially be used to model such sequences. In our study, an empirical method to estimate the distribution of sequence reads along the genome sequence is used. Due to the relative low number of reads starting at particular positions, the empirically estimated read distribution may not be accurate, resulting in less reliable estimated p-values for each pattern. Several investigators studied the distribution of sequence reads from NGS technologies based on the sequence content surrounding a specific location (Hansen, et al. 2010,

Li et al., 2010) and showed that the local sequence content can predict the read distribution well. These results can potentially be used to model the read distribution along the genome sequence. The effect of such dependency on the distribution of the number of occurrences of patterns needs further study. We also assumed that the fragments from NGS are of the same length. In reality, their length can vary and follow some distribution. This is another topic for future studies.

ACKNOWLEDGMENTS

Z.Y.Z. was supported by NSFC 11071146 and Graduate Independent Innovation Foundation of Shandong University (GIIFSDU). G.R. was supported in part by the Institute for Mathematical Sciences of the National University of Singapore and US R21AG032743. M.S.W. was supported by NIH 1R21HG006199. Y.H.L. was supported by China NSFC 11071146. F.S. was supported by NIH P50 HG 002790 and 1R21HG006199; NSF DMS-1043075 and OCE 1136818; and NSFC 60805010.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Campbell, A., Mrázek, J., and Karlin, S. 1999. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* 96, 9184–9189.

Chen, L., and Shao, Q.-M. 2004. Normal approximation under local dependence. *Ann. Probabil.* 32, 1985–2028.

Dalevi, D., Dubhashi, D., and Hermansson, M. 2006. Bayesian classifiers for detecting HGT using fixed and variable order markov models of genomic signatures. *Bioinformatics* 22, 517–522.

Dufraigne, C., Fertil, B., Lespinats, S., et al. 2005. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res.* 33, e6.

Guibas, L., and Odlyzko, A. 1981. Periods in strings. *J. Combin. Theory Ser. A* 30, 19–42.

Hansen, K., Brenner, S., and Dudoit, S. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* 38, e131.

Huang, H. 2002. Error bounds on multivariate normal approximations for word count statistics. *Adv. Appl. Probabil.* 34, 559–586.

Jun, S., Sims, G., Wu, G., et al. S. 2010. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: an alignment-free method with optimal feature resolution. *Proc. Natl. Acad. Sci. USA* 107, 133–138.

Karlin, S., and Mrázek, J. 1997. Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* 94, 10227–10232.

Kleffe, J., and Langbecker, U. 1990. Exact computation of pattern probabilities in random sequences generated by Markov chains. *Comput. Appl. Biosci.* 6, 347–353.

Lander, E., and Waterman, M. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2, 231–239.

Li, J., Jiang, H., and Wong, W. 2010. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol.* 11, R50.

Lothaire, M. 1983. *Algebraic* Combinatorics on Words (*Encyclopedia of Mathematics and Its Applications)*. Cambridge University Press, New York.

Maclean, D., Jones, J., and Studholme, D. 2009. Application of ''next-generation'' sequencing technologies to microbial genetics. *Nat. Rev. Microbiol.* 7, 287–296.

Mardis, E. 2008a. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402.

Mardis, E. 2008b. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141.

McHardy, A., Martín, H., Tsirigos, A., et al. 2006. Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods* 4, 63–72.

Nekrutenko, A., and Li, W. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res.* 10, 1986–1995.

Nuel, G. 2006. Effective p-value computations using Finite Markov Chain Imbedding (FMCI): application to local score and to pattern statistics. *Algorithms Mol. Biol.*, 1 1–14.

Panjer, H. 1981. Recursive evaluation of a family of compound distributions. *Astin Bull.* 12, 22–26.

Pape, U.J., Rahmann, S., Sun, F.Z., et al. 2008. Compound poisson approximation of the number of occurrences of a Position Frequency Matrix (PFM) on both strands. *J. Comput. Biol.* 15: 547–564.

Pavesi, G., Mereghetti, P., Mauri, G., et al. 2004. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* 32, W199–W203.

Reinert, G., and Schbath, S. 1998. Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comput. Biol.* 5, 223–253.

Reinert, G., Schbath, S., and Waterman, M. 2000. Probabilistic and statistical properties of words: an overview. *J. Comput. Biol.* 7, 1–46.

Reinert, G., Schbath, S., and Waterman, M. 2005. Statistics on words with applications to biological sequences, 268–346. *In*: *Applied Combinatorics on Words. (Encyclopedia of Mathemics and Its Applications.* Cambridge University Press, New York.

Robin, S., Rodolphe, F., and Schbath, S. 2005. *DNA, Words and Models: Statistics of Exceptional Words*. Cambridge University Press, New York.

Roquain, E., and Schbath, S. 2007. Improved compound Poisson approximation for the number of occurrences of any rare word family in a stationary Markov chain. *Adv. Appl. Probabil.* 39, 128–140.

Schbath, S. 1995. Compound Poisson approximation of word counts in DNA sequences. *ESAIM Probabil. Stat.* 1, 1–16.

Schbath, S. 2000. An overview on the distribution of word counts in Markov chains. *J. Comput. Biol.* 7, 193–201.

Schbath, S., and Robin, S. 2009. How can pattern statistics be useful for DNA motif discovery?, 319–350. In: *Scan Statistics: Methods and Applications.* (Statistics for Industry and Technology). Birkhäuser, Boston.

Shan G., and Zheng, W.M. 2009. Counting of oligomers in sequences generated by Markov chains for DNA motif discovery. *J. Bioinform. Comput. Biol.* 7, 39–54.

Sims, G., Jun, S., Wu, G., et al. 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. USA* 106, 2677–2682.

Song, K., Ren, J., Zhai, Z.Y., et al. 2012. Alignment-free sequence comparison based on next generation sequencing reads. *Proc. RECOMB 2012* 272–285.

Uberbacher, E., and Mural, R. 1991. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* 88, 11261–11265.

Valouev, A., Johnson, D., Sundquist, A., et al. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* 5, 829–834.

Waterman, M. 1995. *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman & Hall, London.

Willmot, G., and Panjer, H. 1987. Difference equation approaches in evaluation of compound distributions. *Insurance Math. Econ.* 6, 43–56.

Wu, G., Jun, S., Sims, G., et al. 2009. Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. *Proc. Natl. Acad. Sci. USA* 106, 12826–12831.

Zhai, Z.Y., Ku, S., Luan, Y.H., et al. 2010. The power of detecting enriched patterns: an HMM approach. *J. Comput. Biol.* 17, 581–592.

Zhang, Z.D., Rozowsky, J., Snyder, et al. 2008. Modeling ChIP sequencing in silico with applications. *PLoS Comput. Biol.* 4, e1000158.

Address correspondence to:
Dr. Yihui Luan
School of Mathematics
Shandong University
Jinan, Shandong, P.R. China

E-mail: yhluan@sdu.edu.cn

or

Dr. Fengzhu Sun
Molecular and Computational Biology
University of Southern California
1150 Childs Way, RRI201
Los Angeles, CA 90089

E-mail: fsun@usc.edu